Uniform Crime Reporting (UCR) Program Data: A Practitioner's Guide

Jacob Kaplan

2021-03-19

Contents

Preface

If you've read an article about crime or arrests in the United States in the last half century, in most cases it was referring to the FBI's Uniform Crime Reporting Program Data, otherwise known as UCR data. UCR data is, with the exception of the more detailed data that only covers murders, a *monthly number of crimes or arrests reported to a single police agency* which is then gathered by the FBI into one file that includes all reporting agencies. Think of your home town. This data will tell you how many crimes were reported for a small number of crimes or how many people (broken down by age, sex, and race) were arrested for a (larger) set of crimes in that city (if the city has multiple police agencies, it will use the agency which is the primary agency on the case, usually the local police department) in a given month. This is a very broad measure of crime, and its uses in research - or understanding crime at all - is fairly limited. Yet is has become - and will likely remain among researchers for at least the next decade - the most important crime data in the United States.

UCR data is important for three reasons:

1. The definitions are standard and all agencies follow them so you can compare across agencies and over time
2. The data is available since 1960 so there is a long period of available data
3. The data is available for most of the 18,000 police agencies in the United States so you can compare across agencies

For most of this book we'll be discussing the caveats of the above reasons - or, more directly, why these assumptions are wrong - but these are the reasons why the data is so influential.

Motivation

By the end of each chapter you should have a firm grasp on the dataset the covered and how to use it properly. However, this book can't possibly cover every potential use case for the data so make sure to carefully examine the data for your own particular use. This benefits you because you'll know your data better and become a better research because of it. This benefits me because it'll increase the quality of research in my field.

I get a lot of emails from people asking questions about this data so part of my motivation for writing this book is to create a single place that answers as many questions as I can about the data. Again, this is among the most commonly used crime datasets and there are still many current papers published with incorrect information about the data (including such simple aspects like what geographic unit data is in and what time unit it is in). So hopefully this book will decrease the amount of misconceptions about this data, increasing overall research quality.[1]

Structure of the book

This remainder of this book will be divided into eight chapters: an intro chapter briefly summarizing each dataset and going over overall issues with UCR data, six chapters each covering one of the six UCR datasets, and a final one covering county-level data, a highly flawed but common use of the UCR data. Each chapter will follow the same format: we'll start with a history of the data such as when it first because available and important changes in definitions or variables.

Next we'll discuss what the data looks like initially when you get it from the FBI - literally what it looks like in it's fixed-width ASCII format you get from the FBI and what it looks like (what each row and column mean) once it's turned into a useful format that can be read into modern software like R and Stata.[2] For most of the datasets this is a minor process but for data like

[1] Ideally, this will also decrease the number of emails I receive.

[2] To look at the data in its fixed-width ASCII format I'll use the program Notepad which can open up text files like ASCII files. For looking at the machine-readable format I'll use R since I think it's a bit better looking than viewing it in Stata. In both cases,

the arrest or homicide datasets, the conversion process is harder - and this can actually lead to changes in the resulting data. For example, in an old version of the arrest data that I released, I aggregated certain arrestee ages together since my laptop at the time couldn't handle converting data from ASCII to R and Stata without aggregating the ages (more age groups means more columns which means more computer memory needed). So anyone using that data would have less detailed data than the current dataset.

Understanding how the data moves from its rawest form (which in this case is after cleaning by the FBI) is important for being able to truly understand the data and its caveats. However, this is a fairly technical part of each chapter so feel free to skip it. Next in each chapter, we'll cover the variables included in the data and how to use them properly (including not using them at all) - this will be the bulk of each chapter. We'll end each chapter by briefly summarizing the data, how and when it's useful, and - most importantly - when you shouldn't use it.

Since manuals are boring, I'll try to include graphs and images to try to alleviate the boredom. That said, I don't think it's possible to make it too fun so sorry in advanced. This book is a mix of facts about the data, such as how many years are available, and my opinions about it, such as whether it is reliable. In cases of facts I'll just say a statement - e.g. "the offenses data is available since 1960". In cases of opinion I'll temper the statement by saying something like "in my opinion…"

Citing this book

If this book was useful in your research, please cite it. To cite this book, please use the below citation:

Kaplan J (2021). *Uniform Crime Reporting (UCR) Program Data: A Practitioner's Guide*. https://github.com/jacobkap/ucrbook.

BibTeX format:

---

these will be images included in the chapter - you won't need to follow along or use either program.

Sources of UCR data

There are a few different sources of UCR data available today. First, and probably most commonly used, is the data put together by the National Archive of Criminal Justice Data (NACJD)). This a team of out of the University of Michigan who manages a huge number of criminal justice datasets and makes them available to the public. If you have any questions about crime data - UCR or other data - I highly recommend you reach out to them for answers. They have a collection of data and excellent documentation available for UCR data available on their site here. One limitation to their data, however, is that each year of data is available as an individual file meaning that you'll need to concatenate each year together into a single file. Some years also have different column names (generally minor changes like spelling robbery "rob" one year and "robb" the next) which requires more work to standardize before you could concatenate. They also only have data through 2016 which means that the most recent years (UCR data is available through 2019) of data are (as of this writing) unavailable.

Next, and most usable for the general public but limited to researchers, is the FBI's official website Crime Data Explorer. On this site you can chose an agency and see annual crime data (remember, UCR data is monthly so this isn't as detailed as it can be) for certain crimes. This is okay for the general public but only provides a fraction of the data available in the actual data so is really not good for researchers.

Finally, I have my own collection of UCR data available publicly on openICPSR, a site which allows people to submit their data for public access. For each of these datasets I've taken the raw data from the FBI (for early years of homicide data this is actually from NACJD since the FBI's raw data is wrong and can't be parsed. For later years of homicide data this is from the FBI's raw data.) and read it into R. Since the data is only available from the FBI as fixed-width ASCII files, I created a setup file (we'll explain exactly how reading in this kind of data works in the next chapter) and read the data and then very lightly cleaned the data (i.e. only removing extreme outliers like an agency having millions of arsons in a month). For each of these datasets I detail what I've done to the data and briefly summarize the data (i.e. a very short version of this book) on the data's page on openICPSR. The main advantage is that all my data has standard variable names and column names and, for data that is small

enough, provide the data as a single file that has all years. For large datasets like the arrest data I break it down into parts of the data and not all years in a single file. The downside is that I don't provide documentation other than what's on the openICPSR page and only provide data in R and Stata format. I also have a similar site to the FBI's Crime Data Explorer but with more variables available, that site is available here.

It's worth mentioning a final source of UCR information. This is the annual Crimes in the United States report released by the FBI each year around the start of October.[3] As an example, here is the website for the 2019 report. In this report is summarized data which in most cases estimates missing data and provides information about national and subnational (though rarely city-level) crime data. As with the FBI's site it is only a fraction of the true data available so is not a very useful source of crime data. Still, this is a very common source of information used by researchers.

Where to find the data used in this book

The data I am using in this book is the cleaned (we'll discuss in more detail exactly what I did to clean each dataset in the dataset's chapter but the short answer is that I did very little.) and concatenated data that I put together from the raw data that the FBI releases. That data is available on my website here. I am hosting this book through GitHub which has a maximum file size allowed that is far smaller than these data so you'll need to go to my site to download the data, it's not available through this book's GitHub repo. For some examples I'm using the data before I cleaned it of outliers (as an example of the outliers present before I removed them) so that data is not publicly available.

0.1 NIBRS data

Another source of FBI data, and one sometimes considered part of the UCR data collection, is the National Incident-Based Reporting System (NIBRS)

[3]They also release a report about the first 6-months of the most recent year of data before the October release but this is generally an estimate from a sample of agencies so is far less useful.

data. Like its name implies this is an incident-level dataset which has detailed information about each incident reported to the police, including incident circumstances, and victim and offender information. This is also the data that the FBI has declared will replace UCR data starting in 2021, meaning that they will no longer collect UCR data and only allow agencies to submit NIBRS data. NIBRS data is a complex and highly rich dataset that deserves its own book to really understand, so I will not be discussing it any further in this book.

About the author

Jacob Kaplan holds a PhD and a master's degree in criminology from the University of Pennsylvania and a bachelor's degree in criminal justice from California State University, Sacramento. His research focuses on Crime Prevention Through Environmental Design (CPTED), specifically on the effect of outdoor lighting on crime. He is the author of several R packages that make it easier to work with data, including fastDummies and asciiSetupReader. His website allows easy analysis of crime-related data and he has released over a dozen crime data sets (primarily FBI UCR data) on openICPSR that he has compiled, cleaned, and made available to the public.

For a list of papers he has written (including working papers), please see here.

For a list of data sets he has cleaned, aggregated, and made public, please see here.

For a list of R packages he has created, please see here.